Auditing for Score Inflation Using Self-Monitoring Assessments: Findings from Three Pilot Studies

Auditing for Score Inflation Using Self-Monitoring Assessments: Findings from Three Pilot Studies

Prepublication draft (authors' final version)
To be published in *Educational Assessment*

Daniel Koretz[1]
Jennifer L. Jennings[2]
Hui Leng Ng[1]
Carol Yu[1]
David Braslow[1]
Meredith Langi[1]

[1] Harvard Graduate School of Education
[2] New York University

Revised,
July 11, 2016

Acknowledgements: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305AII0420, and by the Spencer Foundation, through Grants 201100075 and 201200071, to the President and Fellows of Harvard College. The authors also thank the New York State Education Department for providing the data used in this study. The opinions expressed are those of the authors and do not represent views of the Institute, the U.S. Department of Education, the Spencer foundation, or the New York State Education Department or its staff.

Abstract

Research has shown that test-based accountability programs often produce score inflation. Most studies have evaluated inflation by comparing trends on a high-stakes test and a lower-stakes audit test. However, Koretz and Beguin (2010) noted the weaknesses of using external audit tests and suggested instead using self-monitoring assessments (SMAs), which incorporate into high-stakes tests audit items that are not susceptible to test preparation aimed at more predictable items. This paper reports the results of the first three trials of the SMA approach, evaluating whether SMAs can detect inflation in a context in which it has been demonstrated to exist. The studies were conducted with the New York State mathematics tests in grades 4, 7, and 8 in 2011 and 2012. Despite a severe conservative bias created by numerous aspects of the study designs, we found that the audit component functioned as expected in many of the trials. The difference in performance between nonaudit and audit items was associated with factors that earlier research showed to be related to test preparation and score inflation, such as "bubble-student" status (scoring just below the Proficient cut in the previous year) and school poverty. However, a number of trials yielded null findings. These findings underscore the need for additional research investigating the optimal characteristics of audit items.

A substantial body of research has shown that educators' instructional responses to test-based accountability programs often produce score inflation, that is, gains in scores substantially higher than warranted by the learning that they are intended to represent (e.g., Jacob, 2007; Hamilton et al., 2007; Klein, Hamilton, McCaffrey, & Stecher, 2000; Koretz & Barron, 1998; Stecher, 2002).[1]

Most studies of score inflation have evaluated the consistency of gains between a high-stakes test and a lower-stakes audit test, most often the National Assessment of Educational Progress (NAEP). The disadvantages of this approach, including concerns about audit tests' substantive appropriateness, their uneven availability across grades and subjects, and the potential for variation in student motivation across tests, have been well documented (Koretz & Beguin, 2010).

As a way to avoid these limitations, Koretz & Beguin (2010) suggested *self-monitoring assessments* (SMAs). SMAs incorporate audit components into the high-stakes test itself, using differences in performance between these audit components and routine operational items as a measure of score inflation. In principle, SMAs could eliminate many of the limitations of using a separate test for detecting score inflation.

This paper presents the results of the first three trials of the SMA approach, all conducted in the context of New York State's mathematics testing program in grades 4, 7, and 8. These studies were conducted solely to evaluate whether it is feasible to construct SMAs that are sensitive to score inflation. Severe inflation in New York's

[1] Score inflation has typically been operationalized as the divergence in trends between scores on a high-stakes test and on a lower-stakes audit test designed to support similar inferences, using either identical students or randomly equivalent groups. For a discussion of methods for validating scores under high-stakes conditions, see Koretz & Hamilton, 2006.

mathematics scores was documented and widely acknowledged because scores on the state's tests had been increasing far faster than those on the National Assessment of Educational Progress (NAEP). For example, the increase in mean scores on the state's eighth-grade mathematics test during the first three years of the testing program was nearly seven times as large as the increase in the state's public-school NAEP scores over the four years spanning that period. This led both the Commissioner and the Chancellor of the Board of Regents to state publicly that performance on the state's tests was misleading and to make a public commitment to make the tests less predictable (Steiner, 2009; Tisch, 2009). It also led them to make a commitment to research on score inflation, including these studies. Given the existing evidence of inflation, our goal in these studies was not to document or quantify inflation but to investigate whether an SMA design is capable of detecting it.

The three studies include numerous replications of our analysis, each of which is low in statistical power and provides a very conservative evaluation of the SMA design for other reasons as well, as discussed below. However, taken together, these multiple replications provide an informative first trial of the SMA approach.

Background

Variations in Score Inflation and Test Preparation

Studies in a variety of contexts have confirmed that test-based accountability in K-12 education often leads to score inflation (e.g., Fuller, Gesicki, Kang, & Wright, 2006; Haladyna, Nolan, & Haas, 1991; Hambleton, et al., 1995; Haney, 2000; Ho, 2007, 2009; Ho & Haertel, 2006; Jacob, 2005, 2007; Jennings & Bearak, 2015; Klein, Hamilton, McCaffrey, & Stecher, 2000; Koretz & Barron, 1998; Koretz, Linn, Dunbar &

Shepard, 1991; Linn, Graue, & Sanders, 1990). Moreover, the bias is often very large. For example, some studies have found gains on state tests two to six times as large as those on NAEP (e.g., Jacob, 2007; Klein, Hamilton, McCaffrey, and Stecher, 2000; Koretz and Barron, 1998). Hambleton et al. (1995) found that gains of roughly three-fourths of a standard deviation on a state's high-stakes fourth-grade reading test were accompanied by no gain whatever on NAEP. A number of studies have estimated smaller effects of coaching for the SAT, often in the range of 0.1-0.2 standard deviation on the mathematics test (e.g., Briggs, 2009; Domingue & Briggs, 2009; Powers & Rock, 1999). However, these studies reflect a different process than test prep in K-12 schools and are methodologically weaker; while most studies of K-12 score inflation rely on comparisons of identical or randomly equivalent groups, studies of SAT coaching rely on covariate-adjustment or propensity-score matching in an attempt to remove differences between coached and uncoached students.

A number of studies have examined forms of test preparation that may inflate scores (e.g., Luna & Turner, 2001; Pedulla et al., 2003; Shepard & Dougherty, 1991; Smith, 1991; Smith & Rottenberg, 1991; Stecher, 2002; Stecher & Barron, 2001). Leaving aside cheating, these entail focusing on the specifics of the test, inflating scores by undermining the test's representation of the domain.

Although most studies of score inflation and test preparation have examined only trends in the student population as a whole, a growing body of research has documented variations in both test preparation and inflation across types of students and schools. Several studies show that the relative gains made by low-income and black or Hispanic students relative to white students on state tests are not matched on audit tests (Klein et

al., 2000; Jacob, 2007; Ho & Haertel, 2006). These findings are consistent with research demonstrating that teachers in schools with student populations that are disproportionately poor or non-Asian minority are likely to focus more than others on assigning drills of test-style items and teaching test-taking strategies (Diamond & Spillane, 2004; Firestone, Camilli, Yurecko, Monfils, & Mayrowetz, 2000; Herman & Golan, 1993; Jacob, Stone, & Roderick, 2004; Lipman, 2002; Madaus et al., 1992; McNeil & Valenzuela, 2001; Urdan & Paris, 1994).

Two studies have directly examined variation in performance on predictable items across student and school populations. Shen (2008) examined differences in the performance of items that were more or less "teachable." She showed that schools had greater improvements over time in performance on more teachable items and that this trend was more pronounced in disadvantaged schools. Similarly, Jennings, Bearak, and Koretz (2011) demonstrated that reductions in the racial achievement gap in New York state mathematics tests were driven by black and Hispanic students' improvement on items testing predictably-assessed standards. These limited studies suggest that variations in test preparation tend to produce the greatest score inflation for the most vulnerable students.

Another documented pattern is the reallocation of instructional resources within schools toward students whose anticipated scores are just under the proficiency cut score that counts most for accountability ("bubble students") and away from students whose anticipated scores are comfortably above or below proficient, a practice described as educational triage (Booher-Jennings, 2005; Gillborn & Youdell, 2000; Neal & Schanzenbach, 2010; Stecher et al., 2008). A number of quantitative studies have

established that "bubble students" appear to make more progress on high-stakes tests when proficiency-based accountability systems are in place (Neal & Schanzenbach, 2010; Reback, 2008), but similar differential progress is not evident on low-stakes tests (Jennings & Sohn, 2014).

Our evaluation of our SMA trials uses these variations as the basis for our analytical strategy. As explained in more detail below, given our design constraints, simple differences in performance between audit items and other items are not directly interpretable. Therefore, we use a difference-in-differences strategy, examining the covariance of the difference between audit and nonaudit items with variables previously found to be correlated with test preparation and score inflation: bubble-student status and student demographics. We consider the bubble-student variable to be the most important of our predictors because the link between bubble status and the incentives facing teachers is direct. In contrast, this incentive is less direct for student demographics, which reflects the associations of these variables with both scores and teacher and school characteristics.

Educators' Responses to Predictable Sampling and the Design of Self-Monitoring Assessments

The opportunity to inflate scores arises because of predictable patterns in tests over time. These patterns can include the amount of emphasis given to different content (including predictable omissions) and the ways in which content is presented. Some of the predictability is intentional (e.g., to keep forms sufficiently similar to permit reasonably error-free linking), while some is unintended. When teachers focus on the predictable specifics of the test, they can undermine its representation of the domain,

hence inflating scores. Test preparation materials often draw users' attention to these predictable patterns and provide strategies for performing well on predictable items (e.g., Rubenstein, 2000).

Focusing preparation on predictable patterns can take two forms, which Koretz & Hamilton (2006) labeled *reallocation* and *coaching*. Both have been documented in the literature (e.g., Hamilton et al, 2007; Luna & Turner, 2001; Pedulla et al., 2003; Shepard & Dougherty, 1991; Smith, 1991; Smith & Rottenberg, 1991; Stecher, 2002; Stecher & Barron, 2001). The distinction bears on our creation of audit components.

Reallocation refers to better aligning instructional resources, such as time, with the content of the specific test used for accountability. Reallocation can be beneficial, but it can also generate score inflation if it entails omitting or de-emphasizing content that is important for the inference based on scores. Numerous studies have found that many teachers report decreasing their emphasis on elements of the curriculum that are de-emphasized by the test (Pedulla et al., 2003; Stecher, 2002).

Coaching entails focusing on minor, often incidental characteristics of the test that are not tied to the target of inference. These can include both substantively unimportant details of content and aspects of task presentation, e.g., item format, the particular graphics used with mathematics items, and so on. For example, item writers typically use Pythagorean triples in items assessing the Pythagorean Theorem because students are unable to compute square roots by hand. One common form of coaching in response to this is telling students to memorize the two Pythagorean triples that most often appear in test items, 3:4:5 ($3^2 + 4^2 = 5^2$) and 5:12:13 (e.g., Rubinstein, 2000). This allows

students to answer the item correctly without actually learning the theorem or being able to apply it in real life.

Audit items in an SMA should assess material that is relevant to the inference without replicating the predictable patterns that create opportunities for test preparation. To use the example above, if all operational items assessing knowledge of the Pythagorean Theorem make use of common Pythagorean triples, a suitable audit item might be a calculator item with a non-integer solution. Students whose test preparation focused on Pythagorean triples in lieu of appropriate instruction about the theorem would be less likely to answer the audit item correctly than would students receiving appropriate instruction on the Pythagorean Theorem.

Data

The New York State Education Department (NYSED) offered us two opportunities to pilot-test an SMA. Changes in the state's operational testing program provided a third unplanned opportunity.

Study A was implemented in the context of a stand-alone field test administered statewide in June 2011 for purposes of linking and pilot-testing new items. We were permitted to insert audit items into 12-item field-test forms that were administered to random samples of students. Six of our audit items were inserted into random positions in these forms. Sample sizes were small, ranging from 2,087 to 2,552 depending on the form (see Table 1).

In response to limitations of Study A—in particular, very small samples and the possibility of downward biases from weak motivation in the field test context—we conducted Study B in the spring of 2012. In this study, we embedded audit items in all operational mathematics test forms in grades 4, 7, and 8. NYSED matrix samples at the

school level: schools are divided into four representative groups, and every student in a given school receives the same test form. Items for scoring students are common across forms, but field test items are matrix sampled. We were permitted to include a maximum of four items per form, and sets of two audit items were placed randomly in 7-item blocks positioned for field testing by the state's contractor. Therefore, the models in this study included large samples of students but samples of only four audit items per student (Table 1).

Study C stemmed from changes NYSED made to its testing program in response to concerns about inflation. In 2011, the tests were made longer, and they included items assessing standards that had not previously been tested. We used the items assessing previously untested standards as an audit component. This gave us very large samples of students and larger samples of items per student than in either of the other studies (Table 1).

Across the three studies, three grades, and multiple forms per grade (ranging from 1 to 4, depending on the study), these studies yielded 25 samples and hence 25 models and tests of our hypothesis: 8 each in grades 4 and 7 and 9 tests in grade 8. Within a grade, all samples—both within and across studies—were independent and randomly equivalent, other than differences between the 2011 and 2012 cohorts. Thus the multiple models per grade can be treated as replications.

These three studies have only partially overlapping strengths and weaknesses, and for that reason, we stress patterns that are consistent across them. Studies A and B gave us control over the items used as audits and included all four of the item types described below. However, both had very low power: Study B had few audit items per student, and Study A had both few audit items and small samples of students per form. Study C had

both large samples of students and larger samples of audit items, but the items were not of our choosing, and they included only one of the four types described below. Study A posed risks of downward motivational biases, whereas studies B and C did not. Operational scores in Study A showed the substantial skewness and right-censoring that is common with high-stakes tests (Ho & Yu, in press). This was less of an issue in the other studies because of the broadening of the state tests in 2011.

We created our analytic samples using the same approach in all three studies. We dropped schools with fewer than 3 tested students. The design of Study A did not permit imposing a higher minimum count. This is discussed further in the section on sensitivity tests below. We then dropped students missing either an audit or nonaudit test score. Lastly, we excluded a small number of students (less than 1%) with apparently anomalous scores.[2] These decisions left us with approximately 94% of the original sample for Study A, 89% for Study C, and 90% for Study B. These analytic samples differed little from our original data in terms of observed student characteristics. Additionally, the samples across the three studies are very similar in demographic composition (see online supplementary materials, Table S.1).

[2] We also dropped a very small number of students with mismatched form booklets. In addition, we dropped P.S. 184 Shuang Wen School, a public school in New York City with an immersion program in Mandarin Chinese. A large percent of the school's students were Asian, and an extreme value relative to the rest of the schools in our sample inflated coefficients for the school proportion-Asian variable.

Methods

Selection of Audit Items

To guide our development of audit items, we modified the framework offered by Holcombe, Jennings, and Koretz (2013) to describe the stages of sampling and narrowing that are unavoidable in the construction of a test. First, one must decide which elements from the target of inference will be represented in the standards or curriculum. The material sampled by the standards may exclude or deemphasize some content that is important for stakeholders' inferences. Second, the test authors must decide which of the standards will be tested, and among those that will be tested, how frequently they will be tested and how much emphasis they will be given in a typical form. Third, a given standard can be worded in various ways, and the specific wording chosen may further narrow the range of what is tested. Finally, the design of the items testing a standard may further narrow a given item may narrow what is tested. This can be either by selection of content from within the range implied by the standard or by the choice of representations of this material.

Each of these stages of sampling, if predictable, affords opportunities to focus on the specifics of the test and thereby undermine representation of the target and inflate scores. The principle underlying the design of SMAs is to administer items that are relevant to the inference but that do not share this predictable narrowing.

Using this framework, we identified opportunities for narrowed instruction in the New York tests. The first stage identified by Holcombe et al. (2013), narrowing from the target to standards, is unavoidably controversial, but as an approximation, we treated the NAEP frameworks and MCAS standards (in grade 7, for which there are no NAEP

frameworks) as operationalizations of the domain. We mapped the New York standards to both and found that the span of the New York State standards was substantially narrower than either. After identifying important material included in the NAEP or MCAS frameworks but not in the New York standards, we selected released items from to represent a small amount of this content. We label these as "not in standards" (NIS) items. NIS items were not used in all of our forms. Because NIS items these are likely to be more controversial than the other types, we clearly identify below which results include them.

To address the subsequent stages of sampling, that is, those that take the New York standards as a given, we examined the wording of each New York standard to see what skills the specific wording includes and excludes. We then used item maps to establish the frequency with which standards had been tested in previous years. Finally, we arrayed all operational items for all years from 2006 to 2010 chronologically within standards to examine similarities and variations in both content and representation. This work revealed predictable patterns in the tests, such as recurrent omissions, persistent differences in content emphasis, and recurrent patterns in the presentation of material.

The second stage arises when certain standards are consistently omitted. The two principal investigators independently reviewed these omitted standards to evaluate their suitability for inclusion in the audit for Studies A and B. Some standards, such as those requiring a calculator or those requiring "use of physical models to perform operations with polynomials" (8A5), were not possible to test in the context of the existing testing program. We prioritized untested standards that asked for critical thinking and application (e.g., 8A14: "Solve linear inequalities by combining like terms, using the distributive

property, or moving variables to one side of the inequality [include multiplication or division of inequalities by a negative number]"), and excluded those that simply specified mathematical vocabulary (e.g., 8A17, "Define and use correct terminology when referring to function [domain and range]) or those asking for mechanical applications of concepts (e.g.,. 8G16, "Determine the equation of a line given the slope and the y-intercept"). We label items testing the selected previously omitted standards as "untested standards" (US) items. In Study C, the US items were provided by the state's contractor, without our input.

Inclusion and exclusion of standards is an extreme instance of the more general problem of predictable differences in emphasis over time. However, our US items audit only standards never before tested, not those tested infrequently in the past.

The third stage, narrowing by wording of the standards, is particularly important in New York. New York wrote many of the standards in use at that time very narrowly. This becomes clear if one compares these standards to the Massachusetts standards. For example, the Massachusetts 8th measurement standard 8M3 states that students will "Demonstrate an understanding of the concepts and apply formulas and procedures for determining measures, including those of area and perimeter/circumference of parallelograms, trapezoids, and circles. Given the formulas, determine the surface area and volume of rectangular prisms, cylinders, and spheres. Use technology as appropriate." In contrast, many of New York's standards are framed narrowly, and in some cases so narrowly that they come close to implying specific test questions. For example, geometry standard 8G4 requires that students, "Determine angle pair relationships when given two parallel lines cut by a transversal." We endeavored to find

audit items that would broaden the scope of some standards modestly, while retaining the construct across which test users might expect performance to generalize. These are labeled “broadening at the level of standards” (BS) items. However, some New York mathematics standards were so narrowly written that we could not devise a way to do this.

The final stages of sampling narrows material from within a standard as it is worded. For example, a given standard may specify that students learn rotation, transformation, and dilation of a polygon on a coordinate plane, but the test items may include only rotations. The items may also show predictable patterns in either the presentation of material or in the use of irrelevant content details. In extreme cases, these choices may be so similar over time that items are virtual clones of items used in earlier forms of the assessment. The audit items addressing these aspects of narrowing are labeled “broadening at the level of items” (BI) items.

Lacking prior research to guide the development or selection of BI or BS items, we opted to try to avoid confounding the effects of preparation aimed at predictable occurrences with other, irrelevant differences between items. Therefore, differences between BI and BS audit items and nonaudit items were minor. This posed the risk of introducing too little novelty and hence underestimates or spurious null findings. The results below do show that some audit items did not function effectively as audits.

In Studies 1 and 3, we were permitted to administer only multiple-choice audit items, and the number of audit items was constrained by the state and its contractor. In Study C, the selection and number of audit items was determined by the state’s contractor. All but one of the Study C audit items were multiple-choice.

The large majority of our audit items for Studies 1 and 3 were publicly available, previously administered items from large-scale assessments, including the NAEP, TIMSS, and state assessments from Massachusetts and Pennsylvania. We prioritized items that asked students to demonstrate a deeper understanding of a skill or concept rather than procedural items, and did not include items with very high p-values in previous administrations. It was often necessary to modify these items, either for content or to match the four-choice format of the New York tests. In a few instances in which we were unable to find suitable previously administered items, we and colleagues at Cito, a testing firm in the Netherlands, wrote items. In Study C, the audit items were US items included in the operational 2011 forms by the state's testing contractor, without guidance from us.

Allocation of Items to Forms

Studies A and B. We included items of all four types in Studies A and B. Counts are shown in Table 2. Because we wanted to re-evaluate the audit design in the context of an operational assessment, many of the audit items administered in Study A were reused in Study B. All items were reused at the seventh grade, and 14 of 16 were reused at the fourth grade. However, we reused only 9 of 16 audit items in eighth grade because we found that several failed to function well based on results from Study A. Therefore, in Study B, we replaced these items with new audit items, selected from the same public sources or created based on the same principles as for Study A.

In Study A, we selected the nonaudit items for each form from past operational and anchor items in the 2011 Field Test. In each form, we matched each of the audit items that were broadened at the standards or item level (BI and BS items) with a

nonaudit item that closely matched it in content but that differed on one or more predictable characteristics potentially associated with coaching. This matching was intended to guard against differences in difficulty arising from intentional differences in content rather than from test preparation. Such matching was not possible by definition for untested-standards (US) and not-in-standards (NIS) items. Nonetheless, we were able to pair each of the US and NIS audit items with a nonaudit item testing the same content area defined by the learning standard or strand (e.g., Algebra) in the New York standards.

In contrast, in Study B, the entire operational form, minus our audit items, was used as the nonaudit component. This was done to lessen the problem of low reliabilities of difference scores created by the small number of audit items, but it precluded the close matching we carried out in Study A. We did post-hoc matching of items in Study B, as described below.

Study C. The audit items in Study C were almost all items that assessed previously untested standards (US items). These were added without input from us. In each of fourth and eighth grades, we dropped three items that assessed previously untested standards but that were extremely easy (*p*-values ranging from .79 to .96), as it is unlikely that extremely easy items would show enough variation to serve effectively as audit items. We used the remaining items in the operational state test as the nonaudit items.

Variables

Outcome. The dependent variable in all three studies was the simple difference between the raw proportion correct on the nonaudit and audit portions of the test:

$$Y_{is} = p_{is}^{non} - p_{is}^{audit} \quad (1)$$

where *i* indexes individuals and *s* indexes schools. We standardized this difference to mean 0, standard deviation 1.

Predictors. Our predictor variables included a variety of student-level characteristics and school averages of these characteristics.

Demographics. We included student-level dummy variables for African-American, Hispanic and Asian students, leaving white and other-race students as our omitted comparison group.[3] We also included an indicator of low-income status which is defined by NYSED as whether or not a student participates in the free and reduced price lunch program or other economic assistance program.[4]

Bubble status. We also included a binary indicator of bubble status. Lacking teachers' own classification of bubble students, we followed the convention in other research (e.g., Neal and Schanzenbach, 2010) and used prior-year performance as a proxy. We classified as bubble students those whose scores in the previous year's state test were up to three raw score points below the cut score for Level 3 ("Proficient"). A sensitivity analysis comparing alternative definitions of this variable is discussed below.

New York City. New York City, which enrolls over 30 percent of the students in the state, might have distinct patterns of inflation because it differs in its demographic makeup and has a unique school accountability system. In Study C, the main effect for the New York City dummy was negative, while the coefficients for the black dummy and proportion black were positive and larger than elsewhere in the state. However, these findings were not replicated in the other studies. Therefore, we did not report the New

[3] In most cases, parents reported race. When parents did not report race, districts were responsible for assigning classifications.

[4] For detailed information about the criteria for the low-income variable, see University of the State of New York (2011), p. 44.

York City dummy in our primary results and excluded interactions with that dummy in the models reported here.

Analytical approach

Lacking an effective way to link across forms, we conducted our analyses at the level of forms within grade.

The audit test score, p_{is}^{audit} in equation (1), is ideally an unbiased estimator of the student's uninflated achievement, θ:

$$p_{is}^{audit} = \theta_{is} + \epsilon_{is}, \qquad E(\epsilon_{is}) = 0. \tag{2}$$

We assume it is unbiased because the items are novel with respect to the specific content or presentation that facilitates inappropriate test preparation and score inflation. To the extent that this assumption is violated, the result is a downward bias in our estimates of inflation.

p_{is}^{non} in equation (1) is potentially biased by inflation, ζ_{is}, which is expected to vary both within and between schools. In addition, school performance may differ between parts of a test for systematic reasons unrelated to inflation, which we label δ_s:

$$p_{is}^{non} = \theta_{is} + \delta_s + \zeta_{is} + \nu_{is}, \qquad E(\nu_{is}) = 0. \tag{3}$$

Therefore one can re-express the audit measure in equation (1) as:

$$Y_{is} = \delta_s + \zeta_{is} + (\nu_{is} - \epsilon_{is}). \tag{4}$$

However, δ_s is unobserved and cannot be directly estimated because we never observe performance in the first year of the testing program, when ζ_{is} would presumably be zero. Therefore, we cannot ascertain the relative difficulty of audit and nonaudit items in the absence of inflation. Our solution, following the suggestion of Koretz & Beguin (2010), is a difference-in-differences approach intended to remove δ_s from our outcome.

Specifically, we investigated whether the nonaudit-audit difference is systematically larger among students or schools for which score inflation is likely to be greater. For this purpose, we relied on the research noted above that investigated where both test preparation and score inflation tend to be most severe. That is, we examined the extent to which the nonaudit-audit difference covaries with poverty, race/ethnicity, and especially bubble status, at both the student and school level.

This approach assumes that δ_s does not substantially covary with our question predictors. There are several reasons why this assumption appears reasonable. First, BI and BS items differ from corresponding nonaudit items only in incidental characteristics, and in the absence of responses to testing, there is no reason to expect these incidental characteristics to vary systematically with the characteristics of students and schools in our models. Likewise, there is no reason to expect that the contractor's decisions about inclusion and exclusion of standards would be systematically related to the characteristics of students or schools. Moreover, with the exception of NIS items, all audit items measure content that is included in the target of inference and therefore in the content that schools are expected to teach, so a failure of gains to generalize to these items can be considered inflation even if we cannot document that it results from response to testing.

The student-level error in audit scores will be large because of the short length of our audit tests, and the resulting low reliability is exacerbated by our use of a difference score. This creates a severe conservative bias—that is, a large risk of Type II error. This problem is ameliorated somewhat in our estimates of school-level relationships, but nonetheless, the lack of statistical power remains a serious problem. This also increases inconsistency in results across forms.

Screening Items and Forms

Because of the small number of audit items in each form, we evaluated the student-level reliability of both the audit scores and the nonaudit-audit difference in all 25 forms.

Although we could not use simple differences in difficulty as an indicator of inflation, we did use these differences as an additional basis for screening forms. Our logic is that if they are functioning as audits, our BI and BS audit items—which are the only audit items that can be matched to nonaudit items—should be more difficult than matched operational items because we changed very little other than adding something novel in content or representation.

In Study C, we had only untested-standards audit items, so this form of matching was not possible. However, we constructed categories of closely related content standards and looked for nonaudit items that assessed standards from the same categories as those for the audit items. Two authors independently evaluated whether any of the matches were close enough to be potentially problematic. We found no problematic matches.

Finally, because of the skewness of operational raw scores, particularly in Study A, we evaluated the skewness of difference scores for all forms.

Regression Models

Our primary model was a two-level random-intercepts model:

(5)
$$Y_{is} = \beta_{0s} + \mathbf{X\beta_{10}} + \beta_{20}N + \epsilon_{\mathrm{is}}$$
$$\beta_{0s} = \gamma_{00} + \mathbf{Z\gamma_{01}} + \mathrm{u}_{0\mathrm{s}}$$

where **X** is a vector of student-level variables, **Z** is the corresponding vector of school-level means, and N is the NYC dummy. This model appropriately adjusts standard errors

for clustering and also accommodates our hypothesis that variations in inflation-inducing behaviors arise between schools.

We grand-mean centered the student-level predictors. This yields parameter estimates for level-2 variables that are direct estimates of context effects (Raudenbush & Bryk, 2002). That is, the parameter estimates for level-2 variables indicate the extent to which the school means differ systematically by more than the level-1 model would predict. We also grand-mean centered level-2 variables for ease of interpretation.

Results

Screening Items and Forms

Across the samples, the distributions of the nonaudit-audit difference scores (our outcome variable) are approximately normal, although the individual distributions of the audit and nonaudit components are in some instances skewed. The skewness statistics for the distributions of difference scores range from .002 to .203. (Figure S.1 in the online supplementary materials shows the distributions of the outcome in grades 4 and 8 in Study C, which are the most skewed and one of the least skewed, respectively).

The internal consistency reliability of the audit item sets varied markedly, from 0.28 to 0.72, with a mean of 0.52. While item-test and item-rest correlations are difficult to interpret when forms are very short, it appears that the low reliabilities reflect the very short lengths of the audit components rather than problematic items. Applying the Spearman-Brown prophecy formula to data pooled across forms within grade suggested that if the audit forms were lengthened only to 10 items each, the reliabilities would fall within a range from 0.46 to 0.76, with most above 0.6.

The estimated reliabilities of the difference scores we use as our outcome variable are very low, ranging from effectively 0 to 0.18[5] (see online supplementary materials, Table S.2.) This is attributable to the substantial correlations between the nonaudit and audit component scores (ranging from 0.57 to 0.81) as well as the low reliability of the audit scores.

It is important to note that the fourth-grade forms produced much more reliable difference scores than the forms in the other grades. Five of eight forms in grade 4 showed a reliability $\geq .10$, a level reached by only two of the 17 forms in grades 7 and 8.

We report here only results from forms for which the reliability of the difference scores was ≥ 0.05. We assume that results from the forms that did not reach this threshold are too noisy to interpret with confidence, and indeed, these models had few statistically significant findings. Results from excluded forms are provided in online supplementary materials (Tables S.3 and S.4).

Simple differences in difficulty varied substantially among forms but did not result in further elimination of forms. In the fourth grade, both matched-item and all-item comparisons showed the audit items to be more difficult than nonaudit items, with *p*-

[5] Conceptually, reliability cannot be negative. In practice, when reliability is very low, one can obtain negative estimates from sampling error. One characteristic of our data increases the probability of negative sample estimates. Using the classical model, the estimated reliability of a difference score will be negative whenever $\left(r_{xx}\sigma_x^2 + r_{yy}\sigma_y^2\right) < 2r_{xy}\sigma_x\sigma_y$, where *x* and *y* are the two tests that are differenced. This inequality is more likely to hold when the test with a larger variance has a considerably lower reliability—precisely what our data produce. Following convention, we set all negative reliability estimates to zero.

value differences ranging from 0.12 to 0.31 (online supplementary Table S.2). In contrast, most of the seventh-grade forms showed very small differences in difficulty, suggesting that our intended audit items did not function as such. Differences in difficulty varied among the eighth-grade forms. However, in most instances, the smallest differences in difficulty occurred in forms already screened out on the basis of the low reliability of difference scores.

One exception warrants explanation because it suggests why our grade 8 audit in Study C failed to provide useful information. Study C had the largest number of audit items, which should increase the reliability of audit scores, and the grade 8 audit scores had the highest reliability of any in that study ($r_{xx'} = .72$). Nonetheless, the reliability of the student-level difference score was near zero ($r_{xx'} = .03$). This reflects the strong correlation between the nonaudit and audit components in this form ($r = .81$) and the small nonaudit-audit difference in scores (0.02). We found that in 2011, the entire distribution of p-values of the operational items shifted downward, and the negative skewness of the raw score distribution was nearly eliminated. This suggests that the grade 8 operational test had been made less predictable and therefore that the audit component may not have been sufficiently different from the nonaudit component to yield noticeable differences in performance.

Regression Results

We present results from the 13 of 25 replications that survived screening for extremely low reliability of the difference score, and we note one pattern from the excluded forms at the end of this section. Because the operational tests differed among grades, we organize our regression results by grade. Because each of the forms within a

grade (even within a single study) were administered to independent, randomly equivalent samples, multiple models per grade serve as replications, increasing confidence and offsetting the low power of each model. Because NIS items are potentially controversial as audits, we identify the replications that include no NIS items by bolding the headings in the tables, and we note these differences in the text. In comparing the results across studies, recall that the problem of low statistical power was most severe in Study A and least severe in Study C.

Because our outcome is $\left(p_{is}^{non} - p_{is}^{audit}\right)$, a positive coefficient indicates support for our hypotheses. For example, a positive sign on the bubble-student variable indicates that performance is weaker on the audit component relative to the nonaudit component for bubble students than for others.

Grade 4. The fourth grade had the largest number of replications (7) that survived screening for low reliability.

The variable most directly related to the pressures that cause inflation, bubble status, showed a substantial relationship to the outcome at the student level in all 7 replications in the fourth grade (Table 4). Across the models, the nonaudit-audit difference was larger for bubble students by 0.16 to 0.38 standard deviations. Moreover, despite very low power, this relationship was statistically significant in all replications and at $p < .001$ in six of the seven models. At the school level, the bubble-student findings were somewhat less consistent. The proportion of bubble students was substantially related to the school mean difference in Studies 2 and 3, and significantly so in three of four models. However, this aggregate relationship was inconsistent in sign and

not significant in the Study A samples, which had the most severe problems of statistical power.

Although it is risky to compare small differences across models, we note that the forms that included the potentially controversial NIS items generally provided *weaker* effects than those that didn't. Five of the fourth-grade replications (Study A, Forms 1 and 2; Study C; and Study B, Forms A and B) contained no NIS audit items, while all of the audit items in the other two forms were NIS. In both Studies A and B, the NIS forms had the smallest coefficients for the student-level bubble variable. In the case of the school-level bubble variable, two of the trials that fail to support the replication are the two NIS forms: Form C in Study B, which yields a nonsignificant coefficient that is much smaller than the other Study B forms, and Form 3 in Study A, which has a nonsignificant negative coefficient.

Fourth-grade results for the low-income variables were weaker and were only moderately consistent with expectations. In Study C, which had the most statistical power, both student-level poverty and the proportion of poor students in the school variables were positively and significantly related to the nonaudit-audit difference, although these estimated relationships were considerably weaker than those of the bubble-status variables. For example, the effect of the student-level poverty dummy in Study C was 0.075 standard deviations, roughly one-third the size of the coefficient for the bubble-status variable. In Study B, results from two forms were consistent with expectations; the exception is again the NIS Form C, which is the form that also did not show a significant relationship with the school proportion of bubble students. The low-

income variables did not predict the outcome significantly in Study A, which had the most severe problem of statistical power.

Results for the race/ethnicity variables were similar to those for low income. Studies B and C showed expected relationships for the student-level Hispanic and Black dummy variables in seven of eight models. However, Study A again did not yield consistent findings for these variables.

Grade 7. Only two seventh-grade forms survived screening for extremely low reliability of the difference scores: Form 3 from Study A, in which 3 of 6 audit items were NIS, and Study C, which had no NIS items. As in grade 4, the student-level bubble-status dummy was substantially related to the outcome; the average bubble-nonbubble difference was approximately 0.25 standard deviations (Table 5). The association with the proportion of bubble students was very large and highly significant in Study C but failed to reach statistical significance in Study A. As in Grade 4, the coefficients were as large or larger when no NIS items were used, but because of the other differences between Studies A and C, this difference are difficult to interpret.

In Study C, which had the most statistical power, all of the other student- and school-level variables conform to expectations: poor students, black students, and Hispanic students all show larger audit-nonaudit differences, and the proportions of all three types of students predict larger school mean outcome differences. As in the fourth grade, these relationships are all smaller than the coefficients of the bubble-status variables. In Study A, however, while the school-level coefficients are in the expected direction, no coefficient other than the student-level bubble-status dummy is statistically significant.

Grade 8. Four forms from the eighth grade survived screening, two each from Study A and Study B. Study C, which generally provided more statistical power, yielded difference scores with a reliability of only 0.03 in this grade, so it was excluded.

The results from the two Study B forms, which contained no NIS items, were only partially consistent with expectations. At the student level, the coefficients of the three demographic dummy variables, black, Hispanic, and low-income, were positive and statistically significant in five of six cases, although the estimates were small, ranging from .04 to .09 standard deviations (Table 5). Most striking, the estimate for the bubble-status dummy was positive and significant in only one of the two forms and was essentially zero in the other. The estimates for the school-level variables in Study B were inconsistent in sign.

The two Study A forms, both of which included NIS audit items (2 and 3 out of 6) yielded no clear patterns.

Results from excluded forms. While we do not report the results for the 12 forms screened out because of very low reliability of the difference score, we noted one pattern in these forms consistent with our expectations. Despite the very low reliability, the coefficients for the bubble-student dummy, the proportion of bubble students, or both were positive and significant in 9 of the 12 excluded forms (see online supplementary materials, Tables S.3 and S.4).

Sensitivity Analyses

We conducted sensitivity tests to evaluate the robustness of our results to possible changes to the minimum count per school and the bandwidth used to define bubble status. Our findings are robust to these changes.

Minimum count per school. Because of the small samples we were allotted in Study A, a threshold higher than $n \geq 3$ would have eliminated too many schools in that study, and we wanted to keep procedures comparable across the three studies. We used data from Study C, which was based on nearly complete census data, to evaluate the effect of replacing this cutoff with $n \geq 10$. This had no appreciable effects on the results.

Bubble status. The definition used in our primary results, three points below proficient in the previous year, classified 6% to 15% of students as bubble students across grades and studies. We replicated our regressions using two alternative definitions: two points below proficient, and from two points below to two points above proficient. These alternative definitions resulted in only modest differences in the proportions of students included and in estimated coefficients and did not produce qualitatively different findings (see online supplementary materials, Table S.5).

Discussion

This paper presents results from the first three trials exploring the feasibility of the SMA approach as a means of detecting score inflation. Our findings should be interpreted in the context of two unavoidable weaknesses. The first was the large risk of Type II errors that stemmed from limits on the number of audit items we could administer, and, in Study A, the small samples of students to whom we could administer each form.

The second limitation was the lack of guidance for the design of our audit components. First, because this was the first attempt to implement SMAs, we had no empirical evidence from prior studies about effective and ineffective approaches to the design of audit items. In addition, we lacked empirical information about the test-

preparation strategies that are widely used in New York. As a result, in the case of items broadened at the level of items or standards, we designed the audit based solely our own *predictions* about coaching strategies based a close review of all past operational test items. Finally, in some cases, our focus on avoiding confounding differences between nonaudit and audit items may have led us to select audit items that were too similar to nonaudit items to effectively identify score inflation. Unfortunately, our data do not permit us to distinguish the possible impact of these factors.

The findings that emerged despite these limitations suggest that the SMA approach has potential as a method for identifying score inflation. Two patterns in particular are worth noting. First, the predictor we consider most important, bubble-student status, yielded quite consistent and in some instances large coefficients, and positive relationships appeared even in the majority of the forms excluded for low reliability. Second, the fourth-grade forms suffered least from the problem of low reliability, and these showed strong and consistent support for the bubble-student hypothesis and a moderate level of support for the hypotheses about demographic factors. We also found mixed support at the 7th and 8th grade levels.

The conventional call for additional research is particularly appropriate here. Since these studies were the first tests of the SMA approach, the need for replications in different contexts is clear. In addition, there is very little research guiding the development of audit items, and we suggest two specific avenues for future research in this area. First, we believe it would be valuable to link the design of items to systematic data on the characteristics of preparation for the operational test in question, rather than relying solely on the characteristics of previously administered items. This information

might be gathered by means of surveys, observational or video studies, or examination of test-preparation materials. Second, additional trials of the SMA approach could be structured to provide comparisons among different item designs—for example, to compare the functioning of several different audit items for a single operational item. This could be accomplished with matrix sampling using paper forms but more easily with computer-based testing.

Finally, the problem of statistical power that we confronted in these studies, which necessitated dropping nearly half of our forms, is itself an important focus for further investigation. Our most serious problems of power arose from the sparse sampling of items, not the sampling of students. Space for audit items will be limited in large-scale testing programs because of limits on total test length. This suggests a need for research on the size of audit components needed for effective use in operational assessments with various test designs.

References

Booher-Jennings, J. (2005). Below the bubble: "Educational Triage" and the Texas Accountability System. *American Educational Research Journal, 42(2), 231-268.*

Briggs, D. C. (2009). *Preparation for College Admission Exams*. Arlington, VA: National Association for College Admission Counseling. Last retrieved July 10, 2016 from http://www.nacacnet.org/research/PublicationsResources/Marketplace/discussion/Pages/TestPreparationDiscussionPaper.aspx.

Diamond, J. B., & Spillane, J. P. (2004). High-stakes accountability in urban elementary schools: Challenging or reproducing inequality? *Teachers College Record, 106*(6), 1145-1176.

Domingue, B. W. & Briggs, D. C. (2009). Using linear regression and propensity score matching to estimate the effect of coaching on the SAT. *Multiple Linear Regression Viewpoints, 35*(1), 12-29.

Figlio, D. and Getzler, L. (2006). Accountability, ability and disability: Gaming the system? In T. Gronberg & D. Jansen (Eds.), *Advances in Applied Microeconomics* (Vol. 14, pp. 35–49). Elsevier.

Firestone, W. A., Camilli, G., Yurecko, M., Monfils, L., & Mayrowetz, D. (2000). State standards, socio-fiscal context and opportunity to learn in New Jersey. *Education Policy Analysis Archives, 8*(35). Retrieved from http://olam.ed.asu.edu/epaa/v8n35.

Fuller, B., Gesicki, K., Kang, E., & Wright, J. (2006). *Is the No Child Left Behind Act working? The reliability of how states track achievement* (Working Paper 06-1). Policy Analysis for California Education, PACE.

Gillborn, D. & Youdell, D. (2000). *Rationing education: policy, practice, reform, and equity*. Buckingham, UK: Open University Press.

Haladyna, T. M., Nolan, S. B., & Haas, N. S. (1991). Raising standardized achievement test scores and the origins of test pollution. *Educational Researcher, 20*(5), 2-7.

Hambleton, R. K., Jaeger, R. M., Koretz, D., Linn, R. L., Millman, J., & Phillips, S. E. (1995). *Review of the measurement quality of the Kentucky Instructional Results Information System, 1991–1994*. Frankfort: Office of Education Accountability, Kentucky General Assembly, June.

Hamilton, L. S., Stecher, B. M., Marsh, J. A., McCombs, J. S., Robyn, A., Russell, J., Naftel, S., and Barney, H. (2007). *Standards-based accountability under No Child Left Behind: Experiences of teachers and administrators in three states*. Santa Monica, CA: RAND Corporation.

Haney, W. (2000). The myth of the Texas miracle in education. *Education Policy Analysis Archives, 8*(41). Retrieved from http://epaa.asu.edu/ojs/article/view/432/828

Herman, J. L., & Golan, S. (1993). The effects of standardized testing on teaching and schools. *Educational Measurement: Issues and Practice, 12*(4), 20-25, 41-42.

Ho, A. D. (2007). Discrepancies between score trends from NAEP and state tests: A scale-invariant perspective. *Educational Measurement: Issues and Practice, 26*(4), 11-20.

Ho, A. D. (2009). A nonparametric framework for comparing trends and gaps across tests. Journal of Educational and Behavioral Statistics, 34(2), 201-228.

Ho, A. D., & Haertel, E. H. (2006). *Metric-Free Measures of Test Score Trends and Gaps with Policy-Relevant Examples* (CSE Report 665). Los Angeles, CA: National Center

for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation, University of California, Los Angeles.

Ho, A. and Yu, C. (in press). Descriptive statistics for modern test score distributions: Skewness, kurtosis, discreteness, and ceiling effects. *Educational and Psychological Measurement*.

Holcombe, R., Jennings, J. L., & Koretz, D. (2013). The roots of score inflation: An examination of opportunities in two states' tests. In G. Sunderman (Ed.), *Charting Reform, Achieving Equity in a Diverse Nation,* 163-189. Greenwich, CT: Information Age Publishing

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools. *Journal of Public Economics, 89*(5-6), 761-796. doi: 10.1016/j.jpubeco.2004.08.004

Jacob, B. (2007). *Test-based accountability and student achievement: An investigation of differential performance on NAEP and state assessments.* Cambridge, MA: National Bureau of Economic Research (Working Paper 12817).

Jacob, R. T., Stone, S., & Roderick, M. (2004). *Ending social promotion: The response of teachers and students*. Chicago, IL: Consortium on Chicago School Research. Retrieved March 29, 2011, from http://www.eric.ed.gov/PDFS/ED483823.pdf

Jennings, J.L & Bearak, J.M. (2014). "Teaching to the test" in the NCLB era: How test predictability affects our understanding of student performance. *Educational Researcher*, 43, 381-389.

Jennings, J.L, Bearak, J.M., & Koretz, D.M. (2011). "Accountability and Racial Inequality in American Education." Paper presented at the Annual Meetings of the American Sociological Association.

Jennings, J.L., & Sohn, H. (2014). Measure for measure: How proficiency-based accountability systems affect inequality in academic achievement. *Sociology of Education,* 87, 125-141.

Klein, S. P., Hamilton, L.S., McCaffrey, D.F., and Stecher, B.M. (2000). *What do test scores in Texas tell us?* Santa Monica, CA: RAND (Issue Paper IP-202). Last accessed from http://www.rand.org/publications/IP/IP202/ on June 4, 2013.

Koretz, D., and Barron, S. I. (1998). *The Validity of Gains on the Kentucky Instructional Results Information System (KIRIS).* MR-1014-EDU, Santa Monica: RAND.

Koretz, D., & Béguin, A. (2010). Self-Monitoring Assessments for Educational Accountability Systems. *Measurement: Interdisciplinary Research & Perspective*, *8*, 92–109. doi:10.1080/15366367.2010.508685

Koretz, D., and Hamilton, L. S. (2006). Testing for accountability in K-12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.), 531-578. Westport, CT: American Council on Education/Praeger.

Koretz, D., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991). The effects of high-stakes testing: Preliminary evidence about generalization across tests, in R. L. Linn (chair), *The Effects of High Stakes Testing*, symposium presented at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, Chicago, April.

Lindquist, E. F. (1951). Preliminary considerations in objective test construction. In E. F. Lindquist (Ed.), *Educational Measurement* (pp. 119-184). Washington, DC: American Council on Education.

Linn, R. L., Graue, M. E., and Sanders, N. M. (1990). Comparing state and district results to national norms: The validity of the claim that "everyone is above average." *Educational Measurement: Issues and Practice, 9* (3), 5-14.

Lipman, P. (2002). Making the global city, making inequality: The political economy and cultural politics of Chicago school policy. *American Educational Research Journal, 39*(2), 379-419.

Luna, C., & Turner, C. L. (2001). The impact of the MCAS: Teachers talk about high-stakes testing. *The English Journal, 91*(1), 79-87.

Madaus, G. F. (1988). The distortion of teaching and testing: High-stakes testing and instruction. *Peabody Journal of Education*, 65(3), 29-46.

Madaus, G. F., West, M. M., Harmon, M. C., Lomax, R. G., Viator, K. A., Mungal, C. F., Butler, P.A., McDomwell, C., Simmons, R., & Sweeney, El. (1992). *The influence of testing on teaching math and science in grades 4-12*. Report by Center for the Study of Testing, Evaluation, and Educational Policy (CSTEEP). MA: CSTEEP

McNeil, L. M. & Valenzuela, A. (2001). The harmful impact of the TAAS system of testing in Texas: Beneath the accountability rhetoric. In M. Kornhaber & G. Orfield (Eds.), *Raising standards or raising barriers? Inequality and high-stakes testing in public education* (pp. 127-150). New York, NY: Century Foundation.

Neal, D. and Schanzenbach, D. (2010). Left behind by design: Proficiency counts and test-based accountability. *Review of Economics & Statistics, 92(2), 263–283*.

New York City Department of Education (2007). Mayor Bloomberg and Chancellor Klein release first-ever public school Progress Reports [Press release]. Retrieved from http://schools.nyc.gov/Offices/mediarelations/NewsandSpeeches/2007-2008/20071105_progress_reports.htm.

Pedulla, J. J., L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers*. Boston, MA: National Board on Educational Testing and Public Policy. Retrieved from http://www.bc.edu/research/nbetpp/statements/nbr2.pdf

Powers, D. & Rock, D. (1999). Effects of coaching on SAT I: Reasoning test scores. *Journal of Educational Measurement, 36* (2), 93-118.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods, second edition*. Thousand Oaks, CA: Sage.

Reback, R. (2008). Teaching to the rating: School accountability and the distribution of student achievement. *Journal of Public Economics, 92*(5-6), 1394-1415.

Rubinstein, J. (2000). *Cracking the MCAS grade 10 math*. New York: Princeton Review Publishing.

Severson, K. (2011). A scandal of cheating, and a fall from grace. *The New York Times*, September 7, p. A16. Last retrieved on June 5, 2013 from http://www.nytimes.com/2011/09/08/us/08hall.html?pagewanted=all&_r=0.

Shen, X. (2008). Do unintended effects of high-stakes testing hit disadvantaged schools harder? (Doctoral dissertation, Stanford University).

Shepard, L. A. (1988). *The harm of measurement-driven instruction*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Shepard, L.A., & Dougherty, K.C. (1991, April). *Effects of high-stakes testing on instruction*. Paper presented at the annual meeting of the American Educational Research Association and National Council on Measurement in Education, Chicago.

Smith, J. L. (1991). Meanings of test preparation. *American Educational Research Journal, 28*(3)*,* 521-542.

Smith, M. L., & Rottenberg, C. (1991). Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and Practice*. *10*(4), 7-11.

Stecher, B. (2002). Consequences of large-scale, high-stakes testing on school and classroom practice. In L. Hamilton, et al., *Test-based Accountability: A Guide for Practitioners and Policymakers*. Santa Monica: RAND. Retrieved from http://www.rand.org/publications/MR/MR1554/MR1554.ch4.pdf..

Stecher, B. M., & Barron, S. I. (2001). Unintended consequences of test-based accountability when testing in "milepost" grades. *Educational Assessment, 7*(4), 259-281.

Stecher, B. M., Epstein, S., Hamilton, L. S., Marsh, J. A., Robyn, A., McCombs, J. S., Russell, J., & Naftel, S. (2008). *Pain and gain: Implementing NCLB in three states, 2004 – 2006*. Santa Monica, CA: RAND. Retrieved from http://www.rand.org/pubs/monographs/2008/RAND_MG784.pdf

Steiner, D. (2009). *Commissioner Steiner's Statement on New York NAEP Performance in Mathematics*. Albany, N.Y.: The State Education Department, Office of Communications (October 14).

Tisch, M. (2009). What the NAEP results mean for New York. Latham, N.Y.: New York School Boards Association (November 9). Last retrieved on June 24, 2012 from http://www.nyssba.org/index.php?src=news&refno=1110&category=On%20Board%20Online%20November%209%202009.

University of the State of New York (2011). *New York State Student Information Repository System (SIRS) manual.* Albany, N. Y.: Author. Last accessed June 24, 2013 from age 244 of http://www.p12.nysed.gov/irs/sirs/archive/2010-11SIRSManual6-2.pdf.

Urdan, T. C., & Paris, S. G. (1994). Teachers' perceptions of standardized achievement tests. *Educational Policy, 8*(2), 137-157.